

On the Sample Complexity of Graphical Model Selection for Non-Stationary Processes

Nguyen Tran Quang and Alexander Jung

Abstract—We formulate and analyze a graphical model selection method for inferring the conditional independence graph of a high-dimensional non-stationary Gaussian random process (time series) from a finite-length observation. The observed process samples are modeled as uncorrelated over time but having different covariance matrices. We characterize the sample complexity of graphical model selection for such processes by analyzing a variant of sparse neighborhood regression. Our results indicate that, similar to the case of i.i.d. samples, accurate GMS is possible even in the high-dimensional regime if the underlying conditional independence graph is sufficiently sparse.

Index Terms—sparsity, graphical model selection, neighborhood regression, high-dimensional statistics.

I. INTRODUCTION

Consider a complex system which is represented by a large number of random variables x_1, \dots, x_p . For the ease of exposition we model those random variables as zero mean jointly Gaussian. We are interested in inferring the conditional independence graph (CIG) of these random variables based on observing samples $\mathbf{x}[n] \in \mathbb{R}^p$, for $n = 0, \dots, N-1$, which are uncorrelated but not identically distributed. The learning method shall cope with the *high-dimensional regime*, where the system size p is (much) larger than the sample size N , i.e., $N \ll p$ [1]–[7]. This problem is relevant, e.g., in the analysis of medical diagnostic data (EEG) [4], climatology [8] and genetics [9].

Contribution: Most existing approaches to graphical model selection (GMS) model the observed data either as i.i.d. or as samples of a stationary random process [3], [6], [10], [11]. By contrast, we model the observed data as an uncorrelated non-stationary process $\mathbf{x}[n]$ having covariance $\mathbf{C}[n]$ which varies with sample index n . Our main conceptual contribution is the formulation of a sparse neighborhood regression GMS method for high-dimensional non-stationary processes. By analyzing this method, we derive upper bounds on the required sample size such that accurate GMS is possible. In particular, our analysis reveals that the crucial parameter determining the required sample size is the minimum average partial correlation between the process components. If this quantity is not too small, accurate GMS is feasible even in the high-dimensional regime where $N \ll p$.

Outline: The remainder of this paper is organized as follows. In Section II we formalize the considered process model and the notion of a CIG. Section III presents a GMS method based on neighborhood regression along with an upper bound on the sample size such that GMS is accurate. The detailed derivation of this bound is in Section IV.

Notation: The maximum (minimum) of two numbers a and b is denoted $a \vee b$ ($a \wedge b$). The set of non-negative real (integer) numbers is denoted \mathbb{R}_+ (\mathbb{Z}_+). Given a p -dimensional process $\mathbf{x}[0], \dots, \mathbf{x}[N-1] \in \mathbb{R}^p$ of length N , we denote its i th scalar component process as $\mathbf{x}_i := (x_i[0], \dots, x_i[N-1])^T \in \mathbb{R}^N$. Given a vector $\mathbf{x} = (x_1, \dots, x_d)^T$, we denote its euclidean and ∞ -norm by $\|\mathbf{x}\|_2 := \sqrt{\sum_i x_i^2}$ and $\|\mathbf{x}\|_\infty := \max_i |x_i|$. The minimum and maximum eigenvalues of a positive semidefinite (psd) matrix \mathbf{C} are denoted $\lambda_{\min}(\mathbf{C})$ and $\lambda_{\max}(\mathbf{C})$, respectively. Given a matrix \mathbf{Q} , we denote its transpose, spectral norm and Frobenius norm by \mathbf{Q}^T , $\|\mathbf{Q}\|_2$ and $\|\mathbf{Q}\|_F$, respectively. It will be handy to define, for a given finite sequence of matrices \mathbf{Q}_l , the block diagonal matrix $\text{blkdiag}\{\mathbf{Q}_l\}$ with l th diagonal block given by \mathbf{Q}_l . The identity matrix of size $d \times d$ is \mathbf{I}_d .

II. PROBLEM FORMULATION

We model the observed data samples $\mathbf{x}[n]$, for $n = 0, \dots, N-1$, as zero-mean Gaussian random vectors, which are uncorrelated, i.e., $E\{\mathbf{x}[n]\mathbf{x}^T[n']\} = \mathbf{0}$ for $n \neq n'$. Thus, the probability distribution of the observed samples is fully specified by the covariance matrices $\mathbf{C}[n] := E\{\mathbf{x}[n]\mathbf{x}^T[n]\}$. We assume the process is suitably scaled such that $\lambda_{\min}(\mathbf{C}[n]) \geq 1$. By contrast to the widely used i.i.d. assumption (where $\mathbf{C}[n] = \mathbf{C}$), we allow the covariance matrix $\mathbf{C}[n]$ to vary with sample index n . However, we impose a smoothness constraint on the dynamics of the covariance. In particular, we assume the covariance matrix $\mathbf{C}[n]$ being constant over evenly spaced length- L blocks of consecutive vector samples. Thus, our sample process model can be summarized as

$$\underbrace{\mathbf{x}[0], \dots, \mathbf{x}[L-1]}_{i.i.d. \sim \mathcal{N}(\mathbf{0}, \mathbf{C}\{b=0\})}, \underbrace{\mathbf{x}[L], \dots, \mathbf{x}[2L-1]}_{i.i.d. \sim \mathcal{N}(\mathbf{0}, \mathbf{C}\{b=1\})}, \dots \quad (1)$$

For ease of exposition and without essential loss of generality, we henceforth assume the sample size N to be an integer multiple of the block length L , i.e.,

$$N = BL \quad (2)$$

with B denoting the number of data blocks.

The model (1) accommodates the case where the observed samples form a stationary process (cf. [10]–[13]) in the following way: Consider a Gaussian zero-mean stationary process $\mathbf{z}[n]$ with auto-covariance function $\mathbf{R}_z[m] := E\{\mathbf{z}[n]\mathbf{z}^T[n-m]\}$ and spectral density matrix $\mathbf{S}_z(\theta) := \sum_m \mathbf{R}_z[m] \exp(-j2\pi\theta m)$ [14]. Let $\mathbf{x}[k] := \sum_{n=0}^{N-1} \mathbf{z}[n] \exp(-j2\pi nk/N)$ denote the discrete Fourier transform (DFT) of the stationary process $\mathbf{z}[n]$. Then,

by well-known properties of the DFT [15], the vectors $\mathbf{x}[k]$ for $k = 0, \dots, N-1$ are approximately uncorrelated Gaussian random vectors with zero mean and covariance matrix $\mathbf{C}[k] \approx \mathbf{S}_z(k/N)$. Moreover, if the effective correlation width W of the process $\mathbf{z}[n]$ is small, i.e., $W \ll N$, the SDM is nearly constant over a frequency interval of length $1/W$. Thus, the DFT vectors $\mathbf{x}[k]$ approximately conform to process model (1) with block length $L = N/W$.

The process model (1) is also useful for the important class of non-stationary processes which are underspread [16]–[19]. A continuous-time random process $\mathbf{z}(t)$ is underspread if its expected ambiguity function (EAF) $\bar{\mathbf{A}}(\tau, \nu) := \int_t \mathbb{E}\{\mathbf{z}(t + \tau/2)\mathbf{z}^T(t - \tau/2)\} \exp(-j2\pi t\nu) dt$ is well concentrated around the origin in the (τ, ν) plane. In particular, if the EAF of $\mathbf{z}(t)$ is supported on the rectangle $[-\tau_0/2, \tau_0/2] \times [-\nu_0/2, \nu_0/2]$, then the process $\mathbf{z}(t)$ is underspread if $\tau_0\nu_0 \ll 1$. One of the most striking properties of an underspread process is that its Wigner-Ville spectrum (which can be loosely interpreted as a time-varying power spectral density) $\bar{\mathbf{W}}(t, f) := \int_{\tau, \nu} \bar{\mathbf{A}}(\tau, \nu) \exp(-2\pi(f\tau - \nu t)) d\tau d\nu$ is approximately constant over a rectangle of area $1/(\tau_0\nu_0)$. Moreover, it can be shown that for a suitably chosen prototype function $g(t)$ (e.g., a Gaussian pulse) and grid constants T, F , the Weyl-Heisenberg set $\{g^{(n,k)}(t) := g(t - nT)e^{-2\pi kFt}\}_{n,k \in \mathbb{Z}}$ [20], yields zero-mean expansion coefficients $\mathbf{x}[n, k] = \int_t \mathbf{z}(t)g^{(n,k)}(t)dt$ which are approximately uncorrelated. The covariance matrix of the vector $\mathbf{x}[(n, k)]$ is approximately given by $\bar{\mathbf{W}}(nT, kF)$. Thus, the vectors $\mathbf{x}[n, k]$ approximately conform to process model (1), with block length $L \approx \frac{1}{TF\tau_0\nu_0}$.

We now define the CIG of a p -dimensional Gaussian process $\mathbf{x}[n] \in \mathbb{R}^p$ conforming to the model (1) as an undirected simple graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with nodes $\mathcal{V} = \{1, \dots, p\}$. Node $i \in \mathcal{V}$ represents the process component $\mathbf{x}_i = (x_i[0], \dots, x_i[N-1])^T$. An edge is absent between nodes i and j , i.e., $\{i, j\} \notin \mathcal{E}$, if the corresponding process components \mathbf{x}_i and \mathbf{x}_j are conditionally independent, given the remaining components $\{\mathbf{x}_r\}_{r \in \mathcal{V} \setminus \{i, j\}}$.

Since we model the process $\mathbf{x}[n]$ as Gaussian (cf. (1)), the conditional independence among the individual process components can be read off conveniently from the inverse covariance (precision) matrices $\mathbf{K}[n] := \mathbf{C}[n]^{-1}$. In particular, \mathbf{x}_i and \mathbf{x}_j are conditionally independent, given $\{\mathbf{x}_r\}_{r \in \mathcal{V} \setminus \{i, j\}}$, if and only if $K_{i,j}[n] = 0$ for all $n = 0, \dots, N-1$ [15, Prop. 1.6.6.]. Thus, we have the following characterization of the CIG \mathcal{G} associated with the sample process $\mathbf{x}[n]$ in (1):

$$\{i, j\} \notin \mathcal{E} \text{ if and only if } K_{i,j}[n] = 0 \text{ for all } n. \quad (3)$$

We highlight the coupling in the CIG characterization (3): An edge is absent, i.e., $\{i, j\} \notin \mathcal{E}$, only if the precision matrix entry $K_{i,j}[n]$ is zero for all $n \in \{0, \dots, N-1\}$.

We will also need a measure for the strength of a connection between process components \mathbf{x}_i and \mathbf{x}_j for $\{i, j\} \in \mathcal{E}$. To this end, we define the *average partial correlation* between \mathbf{x}_i and \mathbf{x}_j as

$$\rho_{i,j} := (1/N) \sum_{n=0}^{N-1} K_{i,j}^2[n] / (K_{i,i}[n]K_{j,j}[n])$$

$$\stackrel{(1)}{=} (1/B) \sum_{b=0}^{B-1} K_{i,j}^2\{b\} / (K_{i,i}\{b\}K_{j,j}\{b\}). \quad (4)$$

By (3) and (4), $\{i, j\} \in \mathcal{E}$ if and only if $\rho_{i,j} \neq 0$.

Accurate estimation of the CIG for finite sample size N (incurring unavoidable sampling noise) is only possible for sufficiently large partial correlations $\rho_{i,j}$ for $\{i, j\} \in \mathcal{E}$.

Assumption 1. *There is a constant $\rho_{\min} > 0$ such that*

$$\rho_{i,j} \geq \rho_{\min} \text{ for any } \{i, j\} \in \mathcal{E}. \quad (5)$$

Moreover, we also assume the CIG underlying $\mathbf{x}[n]$ to be sparse in the sense of having small maximum degree.

Assumption 2. *For some $s < (p/3) \wedge (L/2)$,*

$$|\mathcal{N}(i)| \leq s, \text{ for any node } i \in \mathcal{V}. \quad (6)$$

III. SPARSE NEIGHBORHOOD REGRESSION

The CIG \mathcal{G} of the process $\mathbf{x}[n]$ in (1) is fully specified by the neighborhoods $\mathcal{N}(i) := \{j \in \mathcal{V} \setminus \{i\} : \{i, j\} \in \mathcal{E}\}$, i.e., once we have found all neighborhoods, we can reconstruct the full CIG. In what follows, we focus on the sub-problem of learning the neighborhood $\mathcal{N}(i)$ of an arbitrary but fixed node $i \in \mathcal{V}$. We denote the size of its neighborhood by $s_i := |\mathcal{N}(i)|$.

In view of (1), let us denote for each block $b \in \{0, \dots, B-1\}$ the i th process component as

$$\mathbf{x}_i\{b\} := (x_i[bL], \dots, x_i[(b+1)L-1])^T \in \mathbb{R}^L.$$

According to Lemma IV.3,

$$\mathbf{x}_i\{b\} = \sum_{j \in \mathcal{N}(i)} a_{ij} \mathbf{x}_j\{b\} + \boldsymbol{\varepsilon}_i\{b\}, \quad (7)$$

with the “error term”

$$\boldsymbol{\varepsilon}_i\{b\} \sim \mathcal{N}(\mathbf{0}, (1/K_{i,i}\{b\})\mathbf{I}_L). \quad (8)$$

Moreover, for an index set $\mathcal{T} \subset \mathcal{V} \setminus \{i\}$,

$$\mathbf{x}_i\{b\} = \sum_{j \in \mathcal{T}} a_{ij} \mathbf{x}_j\{b\} + \tilde{\mathbf{x}}_i\{b\} + \boldsymbol{\varepsilon}_i\{b\}, \quad (9)$$

with component $\tilde{\mathbf{x}}_i\{b\} \sim \mathcal{N}(\mathbf{0}, \tilde{\sigma}^2 \mathbf{I}_L)$ being uncorrelated with $\{\mathbf{x}_j\{b\}\}_{j \in \mathcal{T}}$ and $\boldsymbol{\varepsilon}_i\{b\}$. The variance $\tilde{\sigma}^2 = \mathbf{a}^T \tilde{\mathbf{K}}^{-1} \mathbf{a}$ with $\tilde{\mathbf{K}} = ((\mathbf{C}_{\mathcal{N}(i) \cup \mathcal{T}}\{b\})^{-1})_{\mathcal{T}}$ (cf. Lemma IV.3).

In view of the decompositions (7) and (9), it is reasonable to estimate $\mathcal{N}(i)$ via a least-squares search:

$$\hat{\mathcal{N}}(i) := \arg \min_{|\mathcal{T}| \leq s} (1/N) \sum_{b=0}^{B-1} \|\mathbf{P}_{\mathcal{T}\{b\}}^\perp \mathbf{x}_i\{b\}\|_2^2 + \lambda |\mathcal{T}|, \quad (10)$$

with

$$\lambda := \rho_{\min}/2. \quad (11)$$

For a subset $\mathcal{T} = \{i_1, \dots, i_{|\mathcal{T}|}\}$, the matrix $\mathbf{P}_{\mathcal{T}\{b\}} \in \mathbb{R}^{L \times L}$ represents the orthogonal projection on $\text{span}\{\mathbf{x}_j\{b\}\}_{j \in \mathcal{T}}$. The complementary orthogonal projection is

$$\mathbf{P}_{\mathcal{T}\{b\}}^\perp := \mathbf{I}_L - \mathbf{P}_{\mathcal{T}\{b\}}. \quad (12)$$

We can interpret (10) as performing sparse neighborhood regression, since we aim at approximating the i th component \mathbf{x}_i in a sparse manner (by allowing at most s active components) using the remaining process components.

For the analysis of the estimator (10) we require a bound on the eigenvalues of the covariance matrices $\mathbf{C}[n]$.

Assumption 3. For known $\beta \geq 1$, $\lambda_{\max}(\mathbf{C}[n]) \leq \beta$ for all n .

Our main analytical result is an upper bound on the probability of sparse neighborhood regression (10) failing to deliver the true neighborhood. We denote this event by

$$\mathcal{E}_i := \{\mathcal{N}(i) \neq \hat{\mathcal{N}}(i)\}. \quad (13)$$

Theorem III.1. Consider observed samples $\mathbf{x}[n]$ conforming to process model (1) such that Asspt. 1, 2 and 3 are valid. Then, if the minimum average partial correlation satisfies

$$\rho_{\min} \geq 8\beta/(L - 2s) \quad (14)$$

and moreover

$$N(1 - 2s/L) \geq \frac{1200\beta^2}{\rho_{\min}} (4\log(pe) + \log(6/\eta)), \quad (15)$$

the probability of (10) to fail is bounded as

$$\mathbb{P}\{\mathcal{E}_i\} \leq \eta. \quad (16)$$

IV. PROOF OF THE MAIN RESULT

We now provide a detailed proof of Theorem III.1 by analyzing the probability $\mathbb{P}\{\mathcal{E}_i\}$ of the event \mathcal{E}_i (cf. (13)) when (10) fails to deliver the true neighborhood $\mathcal{N}(i)$. Our argument draws heavily on the techniques used in [21].

It will be convenient to introduce the test statistic

$$Z(\mathcal{S}) := (1/N) \sum_{b=0}^{B-1} \|\mathbf{P}_{\mathcal{S}\{b\}}^\perp \mathbf{x}_i\{b\}\|_2^2. \quad (17)$$

For an arbitrary but fixed set \mathcal{T} with $|\mathcal{T}| \leq s$, we define

$$\mathcal{E}_{\mathcal{T}} = \{Z(\mathcal{N}(i)) + \lambda s_i > Z(\mathcal{T}) + \lambda|\mathcal{T}|\}. \quad (18)$$

We will derive an upper bound $M(\ell_1, \ell_2)$ on $\mathbb{P}\{\mathcal{E}_{\mathcal{T}}\}$ which depends on \mathcal{T} only via $\ell_1 = |\mathcal{N}(i) \setminus \mathcal{T}|$ and $\ell_2 = |\mathcal{T} \setminus \mathcal{N}(i)|$. The total number of such subsets $|\mathcal{T}| \leq s$ is

$$N(\ell_1, \ell_2) := \binom{s_i}{\ell_1} \binom{p - s_i}{\ell_2}.$$

Let $[s_i, s] := \{(\ell_1, \ell_2) \in \mathbb{Z}_+^2 : \ell_1 \leq s_i, \ell_2 \leq s, (\ell_1 \vee \ell_2) > 0\}$. By a union bound, $\mathbb{P}\{\mathcal{E}_i\} \leq \sum_{(\ell_1, \ell_2) \in [s_i, s]} N(\ell_1, \ell_2) M(\ell_1, \ell_2)$. Since $\sum_{(\ell_1, \ell_2) \in [s_i, s]} \leq s^2$, we have

$$\log\{\mathbb{P}\{\mathcal{E}_i\}\} \leq \max_{(\ell_1, \ell_2) \in [s_i, s]} \log s^2 N(\ell_1, \ell_2) + \log M(\ell_1, \ell_2). \quad (19)$$

Let us now detail the derivation of the above mentioned upper bound $M(\ell_1, \ell_2)$ on the probability $\mathbb{P}\{\mathcal{E}_{\mathcal{T}}\}$. To this end, in order to make (7) more handy, we stack the vectors $\boldsymbol{\varepsilon}_i\{b\}$ into $\boldsymbol{\varepsilon}_i = (\boldsymbol{\varepsilon}_i\{0\}^T, \dots, \boldsymbol{\varepsilon}_i\{B-1\}^T)^T \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_{\boldsymbol{\varepsilon}_i})$, with

$$\mathbf{C}_{\boldsymbol{\varepsilon}_i} = \text{blkdiag}\{(1/K_{i,i}\{0\})\mathbf{I}_L, \dots, (1/K_{i,i}\{B-1\})\mathbf{I}_L\}.$$

We also need the projection matrix $\mathbf{P}_{\mathcal{T}}^\perp := \text{blkdiag}\{\mathbf{P}_{\mathcal{T}\{b\}}^\perp\}_{b=0}^{B-1}$ (cf. (12)). Rather trivially (cf. (18)),

$$\begin{aligned} \mathcal{E}_{\mathcal{T}} &= \{Z(\mathcal{N}(i)) - (1/N) \|\mathbf{P}_{\mathcal{T}}^\perp \boldsymbol{\varepsilon}_i\|_2^2 \\ &> Z(\mathcal{T}) - (1/N) \|\mathbf{P}_{\mathcal{T}}^\perp \boldsymbol{\varepsilon}_i\|_2^2 + \lambda(|\mathcal{T}| - s_i)\}. \end{aligned} \quad (20)$$

Let us now define, for some number $\delta > 0$ whose precise value to be chosen later, the two events

$$\mathcal{E}_1(\delta) := \{|Z(\mathcal{N}(i)) - (1/N) \|\mathbf{P}_{\mathcal{T}}^\perp \boldsymbol{\varepsilon}_i\|_2^2| \geq \delta\}, \quad (21a)$$

$$\mathcal{E}_2(\delta) := \{Z(\mathcal{T}) - (1/N) \|\mathbf{P}_{\mathcal{T}}^\perp \boldsymbol{\varepsilon}_i\|_2^2 + \lambda(|\mathcal{T}| - s_i) \leq 2\delta\}. \quad (21b)$$

In view of (20), the event $\mathcal{E}_{\mathcal{T}}$ can only occur if at least one of the events $\mathcal{E}_1(\delta)$ or $\mathcal{E}_2(\delta)$ occurs. Therefore, by a union bound,

$$\mathbb{P}\{\mathcal{E}_{\mathcal{T}}\} \leq \mathbb{P}\{\mathcal{E}_1(\delta)\} + \mathbb{P}\{\mathcal{E}_2(\delta)\}. \quad (22)$$

In order to control the event $\mathcal{E}_1(\delta)$, observe

$$\begin{aligned} Z(\mathcal{N}(i)) &\stackrel{(17)}{=} (1/N) \sum_{b=0}^{B-1} \|\mathbf{P}_{\mathcal{N}(i)\{b\}}^\perp \mathbf{x}_i\{b\}\|_2^2 \\ &\stackrel{(7)}{=} (1/N) \sum_{b=0}^{B-1} \left\| \mathbf{P}_{\mathcal{N}(i)\{b\}}^\perp \left(\sum_{i \in \mathcal{N}(i)} a_j \mathbf{x}_j\{b\} + \boldsymbol{\varepsilon}_i\{b\} \right) \right\|_2^2 \\ &= (1/N) \|\mathbf{P}_{\mathcal{N}(i)}^\perp \boldsymbol{\varepsilon}_i\|_2^2. \end{aligned} \quad (23)$$

Hence,

$$\begin{aligned} \mathbb{P}\{\mathcal{E}_1(\delta)\} &\stackrel{(21a)}{=} \mathbb{E}\{\mathbb{P}\{|Z(\mathcal{N}(i)) - (1/N) \|\mathbf{P}_{\mathcal{T}}^\perp \boldsymbol{\varepsilon}_i\|_2^2| \geq \delta \mid \mathbf{x}_{\mathcal{T}}\}\} \\ &\stackrel{(23)}{=} \mathbb{E}\{\mathbb{P}\{(1/N) \|\mathbf{P}_{\mathcal{N}(i)}^\perp \boldsymbol{\varepsilon}_i\|_2^2 - \|\mathbf{P}_{\mathcal{T}}^\perp \boldsymbol{\varepsilon}_i\|_2^2 \geq \delta \mid \mathbf{x}_{\mathcal{T}}\}\} \\ &\stackrel{(8)}{=} \mathbb{E}\{\mathbb{P}\{(1/N) |\mathbf{w}_i^T \mathbf{P} \mathbf{w}_i| \geq \delta \mid \mathbf{x}_{\mathcal{T}}\}\} \end{aligned} \quad (24)$$

with $\mathbf{P} := \text{blkdiag}\{(1/K_{i,i}\{b\})(\mathbf{P}_{\mathcal{N}(i)\{b\}}^\perp - \mathbf{P}_{\mathcal{T}\{b\}}^\perp)\} \in \mathbb{R}^{N \times N}$ and Gaussian random vector $\mathbf{w}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_N)$. By similar arguments as used in [21], one can verify

$$\|\mathbf{P}\|_2 \leq \beta \text{ and } \text{rank}\{\mathbf{P}\} \leq B(\ell_1 + \ell_2) \quad (25)$$

implying, in turn,

$$\|\mathbf{P}\|_{\text{F}}^2 \leq \text{rank}\{\mathbf{P}\} \|\mathbf{P}\|_2^2 \stackrel{(25)}{\leq} B(\ell_1 + \ell_2) \beta^2. \quad (26)$$

Inserting (25), (26) into (59) of Lemma IV.2 yields

$$\mathbb{P}\{\mathcal{E}_1(\delta)\} \leq 2 \exp\left(-\frac{N^2 \delta^2}{8((\ell_1 + \ell_2) B \beta^2 + N \delta \beta)}\right). \quad (27)$$

Thus, whenever

$$N \geq B(\ell_1 + \ell_2) \beta / \delta, \quad (28)$$

we have

$$\mathbb{P}\{\mathcal{E}_1(\delta)\} \leq 2 \exp\left(-\frac{N \delta}{16 \beta}\right). \quad (29)$$

Let us now upper bound the probability of $\mathcal{E}_2(\delta)$ (cf. (21b)). To this end, in view of the decomposition (9), note

$$\mathcal{E}_2(\delta) \subseteq \mathcal{E}_3(\delta) \cup \mathcal{E}_4(\delta) \quad (30)$$

with the events

$$\mathcal{E}_3(\delta) := \{(1/N) \|\mathbf{P}_{\mathcal{T}}^\perp \tilde{\mathbf{x}}_i\|_2^2 + \lambda(|\mathcal{T}| - s_i) \leq 3\delta\} \quad (31)$$

and

$$\mathcal{E}_4(\delta) := \{(2/N) |\tilde{\mathbf{x}}_i^T \mathbf{P}_{\mathcal{T}}^\perp \boldsymbol{\varepsilon}_i| \geq \delta\}. \quad (32)$$

By union bound, (30) implies

$$\mathbb{P}\{\mathcal{E}_2(\delta)\} \leq \mathbb{P}\{\mathcal{E}_3(\delta)\} + \mathbb{P}\{\mathcal{E}_4(\delta)\} \quad (33)$$

such that $\mathbb{P}\{\mathcal{E}_2(\delta)\}$ can be upper bounded by separately

bounding $P\{\mathcal{E}_3(\delta)\}$ and $P\{\mathcal{E}_4(\delta)\}$. Let us define

$$m_3 := E\{(1/N)\|\mathbf{P}_{\mathcal{T}}^\perp \tilde{\mathbf{x}}_i\|_2^2\} + \lambda(|\mathcal{T}| - s_i), \quad (34)$$

with the random vector $\tilde{\mathbf{x}}_i = (\tilde{\mathbf{x}}_i^T\{0\}, \dots, \tilde{\mathbf{x}}_i^T\{B-1\})^T$. According to Corollary IV.4,

$$\tilde{\mathbf{x}}_i\{b\} \sim \mathcal{N}(\mathbf{0}, \tilde{\sigma}_b^2 \mathbf{I}_L) \text{ with } \tilde{\sigma}_b^2 \geq \ell_1 \rho_{\min}/\beta. \quad (35)$$

Therefore,

$$\begin{aligned} m_3 &\stackrel{(34)}{=} (1/N) \text{Tr}\{\mathbf{C}_{\tilde{\mathbf{x}}_i}^{1/2} \mathbf{P}_{\mathcal{T}}^\perp \mathbf{C}_{\tilde{\mathbf{x}}_i}^{1/2}\} + \lambda(|\mathcal{T}| - s_i) \\ &\stackrel{(35)}{\geq} (B(L-\ell_1)/N) \ell_1 \rho_{\min} + \lambda(|\mathcal{T}| - s_i) \\ &\stackrel{(11)}{\geq} (B(L-\ell_1^2/\bar{\ell})/N) \bar{\ell} \rho_{\min}, \end{aligned} \quad (36)$$

with

$$\bar{\ell} := \ell_1 + (|\mathcal{T}| - s_i)/2 = (\ell_1 + \ell_2)/2 \geq (\ell_1 \vee \ell_2)/2. \quad (37)$$

Let us now choose

$$\delta := m_3/4. \quad (38)$$

Observe

$$\begin{aligned} P\{\mathcal{E}_3(\delta)\} &\stackrel{(31)}{=} E\{P\{(1/N)\|\mathbf{P}_{\mathcal{T}}^\perp \tilde{\mathbf{x}}_i\|_2^2 \leq 3\delta \mid \mathbf{x}_{\mathcal{T}}\}\} \\ &\stackrel{(38)}{\leq} E\{P\{|(1/N)\|\mathbf{P}_{\mathcal{T}}^\perp \tilde{\mathbf{x}}_i\|_2^2 - m_3| \geq \delta \mid \mathbf{x}_{\mathcal{T}}\}\}. \end{aligned} \quad (39)$$

Applying Lemma IV.2 to (39) gets us to

$$P\{\mathcal{E}_3(\delta)\} \leq 2 \exp\left(-\frac{(N/\beta)\delta^2}{8(m_3+\delta)}\right) \stackrel{(38)}{=} 2 \exp\left(-\frac{Nm_3}{\beta 160}\right). \quad (40)$$

In order to control $P\{\mathcal{E}_4(\delta)\}$ (cf. (32)), note

$$P\{\mathcal{E}_4\} = E\{P\{\mathbf{w}^T \mathbf{Q} \mathbf{w} \geq \delta \mid \mathbf{x}_{\mathcal{T}}\}\}, \quad (41)$$

with $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_N)$ and the symmetric matrix

$$\mathbf{Q} = \frac{2}{N} \begin{pmatrix} \mathbf{C}_{\tilde{\mathbf{x}}_i}^{1/2} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{\tilde{\mathbf{x}}_i}^{1/2} \end{pmatrix} \begin{pmatrix} \mathbf{0} & \mathbf{P}_{\mathcal{T}}^\perp \\ (\mathbf{P}_{\mathcal{T}}^\perp)^T & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{C}_{\tilde{\mathbf{x}}_i}^{1/2} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{\tilde{\mathbf{x}}_i}^{1/2} \end{pmatrix}.$$

Applying Lemma IV.2 to (41) gets us to

$$\begin{aligned} P\{\mathcal{E}_4(\delta)\} &\leq 2 \exp\left(-\frac{N(\delta/2)^2}{8(\beta m_3 + \beta(\delta/2))}\right) \\ &\stackrel{(38)}{=} 2 \exp\left(-\frac{(N/\beta)m_3}{16 \cdot 36}\right). \end{aligned} \quad (42)$$

For the choice $\delta = m_3/4$ (cf. (38)), the condition (14) implies validity of (28). Therefore, (29) is in force and can be reformulated using (38) as

$$P\{\mathcal{E}_1(\delta)\} \leq 2 \exp\left(-\frac{Nm_3}{64\beta}\right). \quad (43)$$

Combining (43), (42), (40) with (33), (22) and (36) yields

$$\begin{aligned} P\{\mathcal{E}_{\mathcal{T}}\} &\leq 6 \exp(-\rho_{\min} \bar{\ell} B(L-\ell_1^2/\bar{\ell})/(576\beta^2)) \\ &\stackrel{(37)}{\leq} 6 \exp(-(1/2)\rho_{\min}(\ell_1 \vee \ell_2)B(L-2\ell_1)/(576\beta^2)). \end{aligned} \quad (44)$$

We finalize the proof of Theorem III.1, by using the RHS

of (44) as $M(\ell_1, \ell_2)$ in (19). Thus, $P\{\mathcal{E}_i\} \leq \eta$ holds if

$$\max_{(\ell_1, \ell_2) \in [s_i, s]} \log \frac{6s^2 N(\ell_1, \ell_2)}{\eta} - \frac{\rho_{\min}(\ell_1 \vee \ell_2)B(L-2\ell_1)}{2 \cdot 576\beta^2} \leq 0. \quad (45)$$

The validity of (45), in turn, is guaranteed if

$$B(L-2\ell_1) \geq \frac{2 \cdot 576\beta^2}{\rho_{\min}(\ell_1 \vee \ell_2)} (\log s^2 N(\ell_1, \ell_2) + \log(6/\eta)), \quad (46)$$

for all $(\ell_1, \ell_2) \in [s_i, s]$. Since $s \leq p/3$ (cf. Asspt. 2),

$$s^2 N(\ell_1, \ell_2) \leq \left(\frac{p-s_i}{(\ell_1 \vee \ell_2)}\right)^4 \stackrel{(a)}{\leq} \left[\frac{pe}{\ell_1 \vee \ell_2}\right]^{4(\ell_1 \vee \ell_2)} \quad (47)$$

where (a) is due to $\binom{p}{q} \leq \left(\frac{pe}{q}\right)^q$ [21]. Combining (46) and (47), we finally obtain (15) of Theorem III.1.

APPENDIX

In order to make this paper somewhat self-contained, we list here some well-known facts about Gaussian random vectors, which are instrumental for the derivation in Section IV.

Lemma IV.1. For N i.i.d. standard normal variables $z_i \sim \mathcal{N}(0, 1)$, consider the weighted sum of squares

$$y = \sum_{i=1}^N a_i z_i^2 \quad (48)$$

with weight vector $\mathbf{a} = (a_1, \dots, a_N)^T \in \mathbb{R}^N$. We have

$$P\{|y - E\{y\}| \geq \delta\} \leq 2 \exp\left(-\frac{\delta^2/8}{\|\mathbf{a}\|_2^2 + \|\mathbf{a}\|_\infty \delta}\right). \quad (49)$$

Proof: Consider some $\lambda \in [0, 1/(4\|\mathbf{a}\|_\infty)]$, such that

$$E\{\exp(\lambda a_i z_i^2)\} = 1/\sqrt{1-2\lambda a_i}, \quad (50)$$

and hence,

$$\begin{aligned} \log E\{\exp(\lambda a_i(z_i^2 - 1))\} &= (-1/2) \log(1 - 2\lambda a_i) - \lambda a_i \\ &\leq (-1/2) \log(1 - 2\lambda|a_i|) - \lambda|a_i|. \end{aligned} \quad (51)$$

Applying the elementary inequality

$$-\log(1-u) \leq u + \frac{u^2}{2(1-u)}, \quad 0 \leq u \leq 1 \quad (52)$$

to the RHS of (51) yields

$$\log E\{\exp(\lambda a_i(z_i^2 - 1))\} \leq \frac{\lambda^2 a_i^2}{1 - 2\lambda|a_i|} \leq 2\lambda^2 a_i^2. \quad (53)$$

Summing (53) over i yields

$$\log E\{\exp(\lambda(y - E\{y\}))\} \leq 2\lambda^2 \|\mathbf{a}\|_2^2. \quad (54)$$

Now, consider the tail bound

$$\begin{aligned} P\{y - E\{y\} \geq \delta\} &\leq \exp(-\lambda\delta) E\{\exp(\lambda(y - E\{y\}))\} \\ &\stackrel{(54)}{\leq} \exp(-\lambda\delta + 2\lambda^2 \|\mathbf{a}\|_2^2). \end{aligned} \quad (55)$$

Minimizing the RHS of (55) over $\lambda \in [0, 1/(4\|\mathbf{a}\|_\infty)]$,

$$\begin{aligned} P\{y - E\{y\} \geq \delta\} &\leq \exp\left(-(\delta^2/8)((1/\|\mathbf{a}\|_2^2) \wedge (1/(\delta\|\mathbf{a}\|_\infty)))\right) \\ &\stackrel{(a)}{\leq} \exp\left(-\frac{\delta^2/8}{\|\mathbf{a}\|_2^2 + \|\mathbf{a}\|_\infty \delta}\right), \end{aligned} \quad (56)$$

where (a) is due to $(1/x) \wedge (1/y) \geq 1/(x+y)$ for $x, y \in \mathbb{R}_+$.

Similar to (55), one can also verify

$$P\{y - E\{y\} \leq -\delta\} \leq \exp\left(-\frac{\delta^2/8}{\|\mathbf{a}\|_2^2 + \|\mathbf{a}\|_\infty \delta}\right). \quad (57)$$

Adding (56) and (57) yields (49) by union bound. ■

Lemma IV.2. Consider a Gaussian random vector $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and a symmetric matrix $\mathbf{Q} \in \mathbb{R}^{N \times N}$, i.e., $\mathbf{Q}^T = \mathbf{Q}$. Then, the quadratic form

$$y = \mathbf{z}^T \mathbf{Q} \mathbf{z} \quad (58)$$

satisfies

$$P\{|y - E\{y\}| \geq \delta\} \leq 2 \exp\left(-\frac{\delta^2}{8(\|\mathbf{Q}\|_F^2 + \|\mathbf{Q}\|_2 \delta)}\right) \quad (59)$$

with $E\{y\} = \text{Tr}\{\mathbf{Q}\}$.

Proof: The spectral decomposition of \mathbf{Q} yields [22]

$$\mathbf{Q} = \sum_{i=1}^N \lambda_i \mathbf{u}_i \mathbf{u}_i^T \quad (60)$$

with eigenvalues $\lambda_i \in \mathbb{R}$ and orthonormal eigenvectors $\{\mathbf{u}_i\}_{i=1}^N$. Inserting (60) into (58) yields

$$y = \sum_{i=1}^N \lambda_i (\mathbf{u}_i^T \mathbf{z})^2. \quad (61)$$

Note that $\|\mathbf{Q}\|_2 = \max_{i=1}^N |\lambda_i|$ and $\|\mathbf{Q}\|_F^2 = \sum_{i=1}^N \lambda_i^2$. Since the random variables $\{\mathbf{u}_i^T \mathbf{z}\}_{i=1}^N$ are i.i.d. standard normal, we can apply Lemma IV.1 to (61) yielding (59). ■

Lemma IV.3. Consider a Gaussian random vector $\mathbf{x} = (x_1, \dots, x_p) \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$ with precision matrix $\mathbf{K} = \mathbf{C}^{-1}$. For an arbitrary index set $\mathcal{T} \subseteq \mathcal{V}$, we have

$$x_i = \sum_{j \in \mathcal{T}} c_j x_j + \mathbf{a}^T \tilde{\mathbf{x}} + \varepsilon_i. \quad (62)$$

with fixed weight vector $\mathbf{a} = \{K_{i,j}/K_{i,i}\}_{j \in \mathcal{N}(i) \setminus \mathcal{T}}$ and the random vector $\tilde{\mathbf{x}} \sim \mathcal{N}(\mathbf{0}, \tilde{\mathbf{C}})$, which is independent of $\{x_j\}_{j \in \mathcal{T}}$. The error term ε_i is Gaussian $\sim \mathcal{N}(0, K_{i,i}^{-1})$ and independent of $\tilde{\mathbf{x}}$ and $\{x_j\}_{j \in \mathcal{T}}$. The covariance matrix of $\tilde{\mathbf{x}}$ is given via $\tilde{\mathbf{C}}^{-1} = ((\mathbf{C}_{\mathcal{T} \cup \mathcal{N}(i)})^{-1})_{\mathcal{N}(i) \setminus \mathcal{T}}$.

Corollary IV.4. Consider a vector-valued process $\mathbf{x}[n]$ of the form (1) such that Asspt. 1 and 3 are valid. The i th component $\mathbf{x}_i\{b\}$ can be decomposed as

$$\mathbf{x}_i\{b\} = \sum_{j \in \mathcal{T}} c_j \mathbf{x}_j\{b\} + \tilde{\mathbf{x}}_i\{b\} + \varepsilon_i\{b\}. \quad (63)$$

with $\tilde{\mathbf{x}}_i\{b\} \sim \mathcal{N}(\mathbf{0}, \tilde{\sigma}^2 \mathbf{I}_L)$, which is independent of $\{\mathbf{x}_j\{b\}\}_{j \in \mathcal{T}}$ and has variance

$$\tilde{\sigma}^2 \geq |\mathcal{N}(i) \setminus \mathcal{T}| \rho_{\min} / \beta. \quad (64)$$

The error term $\varepsilon_i\{b\} \sim \mathcal{N}(\mathbf{0}, (1/K_{i,i}\{b\})\mathbf{I}_L)$ is independent of $\tilde{\mathbf{x}}_i\{b\}$ and $\{\mathbf{x}_j\{b\}\}_{j \in \mathcal{T}}$.

REFERENCES

- [1] N. E. Karoui, "Operator norm consistent estimation of large-dimensional sparse covariance matrices," *Ann. Statist.*, vol. 36, no. 6, pp. 2717–2756, 2008.
- [2] N. P. Santhanam and M. J. Wainwright, "Information-theoretic limits of selecting binary graphical models in high dimensions," *IEEE Trans. Inf. Theory*, vol. 58, no. 7, pp. 4117–4134, Jul. 2012.

- [3] P. Ravikumar, M. J. Wainwright, and J. Lafferty, "High-dimensional Ising model selection using ℓ_1 -regularized logistic regression," *Ann. Stat.*, vol. 38, no. 3, pp. 1287–1319, 2010.
- [4] A. Bolstad, B. D. van Veen, and R. Nowak, "Causal network inference via group sparse regularization," *IEEE Trans. Signal Processing*, vol. 59, no. 6, pp. 2628–2641, Jun. 2011.
- [5] J. Bento, M. Ibrahimi, and A. Montanari, "Learning networks of stochastic differential equations," in *Advances in Neural Information Processing Systems 23*, Vancouver, CN, 2010, pp. 172–180.
- [6] N. Meinshausen and P. Bühlmann, "High-dimensional graphs and variable selection with the Lasso," *Ann. Stat.*, vol. 34, no. 3, pp. 1436–1462, 2006.
- [7] J. H. Friedmann, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, vol. 9, no. 3, pp. 432–441, Jul. 2008.
- [8] I. Ebert-Uphoff and Y. Deng, "A new type of climate network based on probabilistic graphical models: Results of boreal winter versus summer," *Geophysical Research Letters*, vol. 39, no. 19, [Online]. Available: <http://dx.doi.org/10.1029/2012GL053269>
- [9] E. Davidson and M. Levin, "Gene regulatory networks," *Proc. Natl. Acad. Sci.*, vol. 102, no. 14, Apr. 2005.
- [10] A. Jung, G. Hannak, and N. Görtz, "Graphical LASSO Based Model Selection for Time Series," to appear in *IEEE Sig. Proc. Letters*, 2015.
- [11] A. Jung, "Learning the conditional independence structure of stationary time series: A multitask learning approach," submitted to *IEEE Trans. Sig. Proc.*, also available at <http://arxiv.org/pdf/1404.1361v3.pdf>, 2014.
- [12] A. Jung, R. Heckel, H. Bölcskei, and F. Hlawatsch, "Compressive nonparametric graphical model selection for time series," in *Proc. IEEE ICASSP-2014*, Florence, Italy, May 2014.
- [13] G. Hannak, A. Jung, and N. Görtz, "On the information-theoretic limits of graphical model selection for Gaussian time series," in *Proc. EUSIPCO 2014*, Lisbon, Portugal, 2014.
- [14] R. Dahlhaus, "Graphical interaction models for multivariate time series," *Metrika*, vol. 51, pp. 151–172, 2000.
- [15] P. J. Brockwell and R. A. Davis, *Time Series: Theory and Methods*. New York: Springer, 1991.
- [16] A. Jung, G. Tauböck, and F. Hlawatsch, "Compressive nonstationary spectral estimation using parsimonious random sampling of the ambiguity function," in *Proc. IEEE-SSP 2009*, Cardiff, Wales, UK, Aug.–Sep. 2009, pp. 642–645.
- [17] —, "Compressive spectral estimation for nonstationary random processes," in *Proc. IEEE ICASSP-2009*, Taipei, Taiwan, R.O.C., April 2009, pp. 3029–3032.
- [18] A. Jung, G. Tauböck, and F. Hlawatsch, "Compressive spectral estimation for nonstationary random processes," *IEEE Trans. Inf. Theory*, vol. 59, no. 5, pp. 3117–3138, May 2013.
- [19] G. Matz and F. Hlawatsch, "Nonstationary spectral analysis based on time-frequency operator symbols and underspread approximations," *IEEE Trans. Inf. Theory*, vol. 52, no. 3, pp. 1067–1086, March 2006.
- [20] G. Durisi, U. G. Schuster, H. Bölcskei, and S. Shamai, "Noncoherent capacity of underspread fading channels," *IEEE Trans. Inf. Theory*, vol. 56, no. 1, pp. 367–395, Jan 2010.
- [21] M. J. Wainwright, "Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting," *IEEE Trans. Inf. Theory*, vol. 55, no. 12, pp. 5728–5741, Dec. 2009.
- [22] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge, UK: Cambridge Univ. Press, 1985.